The R language

Martin Frigaard

2021-08-27

Contents

List of Tables

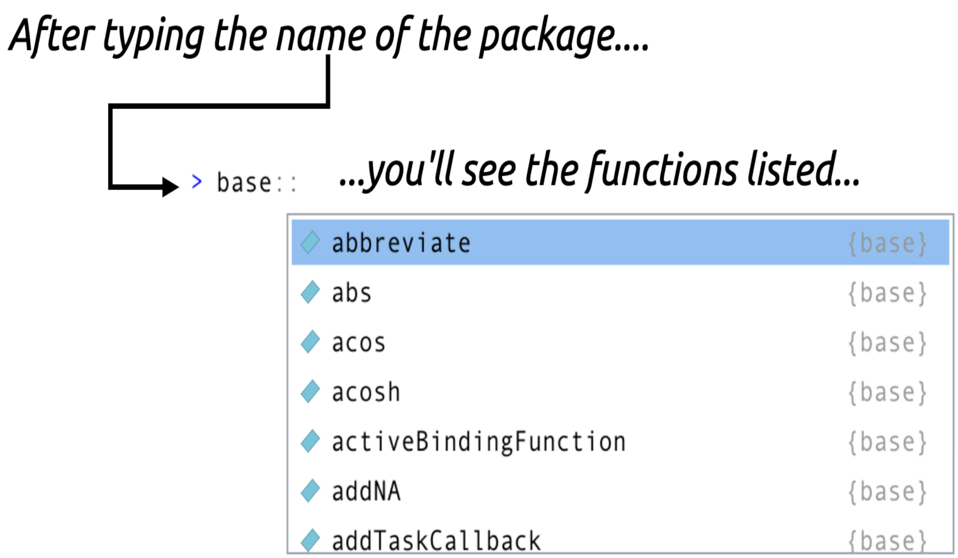
List of Figures

R is an object-oriented, functional programming language for statistical analysis and graphics. R is also a free and open-source software (FOSS) with a massive global community of users and developers who have helped create and maintain tools for data manipulation, graphics, statistics, and machine learning.

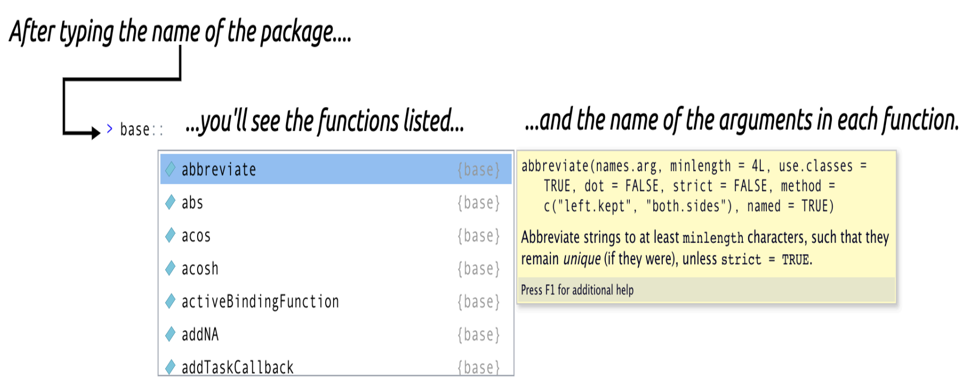
Base R packages

R `packages` are collections of commands for a particular purpose or task. R comes 'out of the box' with a ton of functions for manipulating, analyzing, and visualizing data. Two of the most commonly used standard packages are `base` and `utils`.

You can access any function in a package using the `package::function()` syntax. If you're in RStudio, you can actually see the functions in each package by using the tab-completion feature:



If you hover over the function with your mouse cursor, you'll also see the arguments and documentation for each function.



When you're using the `utils::install.packages()` function, the package files are installed from the Comprehensive R Archive Network, or CRAN. These packages have passed a variety of tests and are generally considered to be more reliable.

User-written packages

Most of the packages we'll be using in this course come from the `tidyverse`, which is a suite of tools pioneered by RStudio's Chief Scientist Hadley Wickham. All packages in the `tidyverse` work well together because they center around a common thread of tidy data.

To install and load the `tidyverse`, we will use the `utils::install.packages()` function to download and installs R packages into a local folder on our computer, and the `base::library()` command loads the packages.

```
install.packages("tidyverse")
library(tidyverse)
```

NOTE: Not all functions return an output. Some functions return messages (or prompts), so be sure to check the help files by using `?install.packages` in the console.

User-written packages can be installed from code repositories like Github.. The R ecosystem has over 10,000 user-written packages available on CRAN.

First you will need to install the `remotes` package from CRAN

```
install.packages("remotes")
```

Second, you load the package with `library()`

```
library(remotes)
```

Finally, we use the `remotes::install_github()` to download and install the `tidyverse` package.

```
remotes::install_github("tidyverse/tidyverse")
library(tidyverse)
```

Note: when installing packages from Github or other repos, you're getting the 'freshest' version, so there might be bugs or errors. If you run into an issue, look for a version of the package on CRAN

Functions and objects

The R language is comprised of functions and objects. R uses functions to perform operations (like `mean()`, `sum()`, `lm()` (for linear model)) on objects (vectors, arrays, matrices, data.frames or lists).

Generally speaking, functions are similar to *verbs*, and objects are more like *nouns*. Functions typically take an object as an `input`, perform an operation on that object, and then return an `output` object.

```
object <- function('input') {

    perform operation(s) on 'input'

    return output
}
# view object
object
```

Creating objects

R comes with a variety of functions for creating objects. We will start with `c()`, which stands for 'combine' or 'concatenate'.

We can print this to the console by supplying the new object and hitting enter/return.

```
x <- c(42, 34, 28, 53, 71, 30, 23, 72, 59, 46, 64, 33, 42, 50, 68)
x
```

```
##  [1] 42 34 28 53 71 30 23 72 59 46 64 33 42 50 68
```

A quick note on printing: notice the preceding `[1]` *in the output. This is not part of the object, it's the line number for the output.*

Now that we have an object in R, what do we do with it? We will start by taking a look at some of it's technical information using `class()` and `str()`

```
class(x)
```

```
## [1] "numeric"
```

The `class()` function tells us `x` is a `numeric` vector. The `str()` function is an abbreviation for 'structure', and it gives us a bit more information.

```
str(x)
```

```
##  num [1:15] 42 34 28 53 71 30 23 72 59 46 ...
```

I recommend using `str()` and `class()` when you're programming in R. Knowing what kind of object you're dealing with will help you determine what you can do with it.

Store and explore

Given the relationship between functions and objects, a common workflow is '*store and explore*, where we create (or import) some data as an object, apply a function to this object, store the output from this function in a new object, and use *another* function to view the result.

Store

Below we create `my_data`, a vector of 12 numbers, and print it to the console.

```
# create data
my_data <- c(49, 147, 74, 90, 7, -79, 190, 49, -123, -325, 143, 232)
my_data
```

```
##  [1]   49  147   74   90    7  -79  190   49 -123 -325  143  232
```

Next, we apply the `sqrt()` function to this object, and store the output from this function in a new object called `my_result`.

```
# apply a function and store in my_result
my_result <- sqrt(my_data)
```

```
## Warning in sqrt(my_data): NaNs produced
```

Warnings and messages

R usually tells us when we've used a function and it's produced a result we might not expect (like missing values). In this case, this missing values aren't important, but it's a good idea to pay attention to any warnings or messages that are printed to the console.

Explore

Now we use the `summary()` function to explore the contents of `my_result`:

```
# explore the result
summary(my_result)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   2.646   7.000   9.487   9.759  12.124  15.232       3
```

This process (applying functions to data objects, storing the results, and using functions to view their contents) makes up the majority of the R workflow.

Operators

Operators are symbols (or collections of symbols) for performing arithmetic (+, `-`, `*`, `/`), comparisons (<, >, =<, =>), and assignment (`<-` and =). These aren't a new kind of object, though (*operators* are also functions! See below).

```
class(`+`)
```

```
## [1] "function"
```

```
class(`<=`)
```

```
## [1] "function"
```

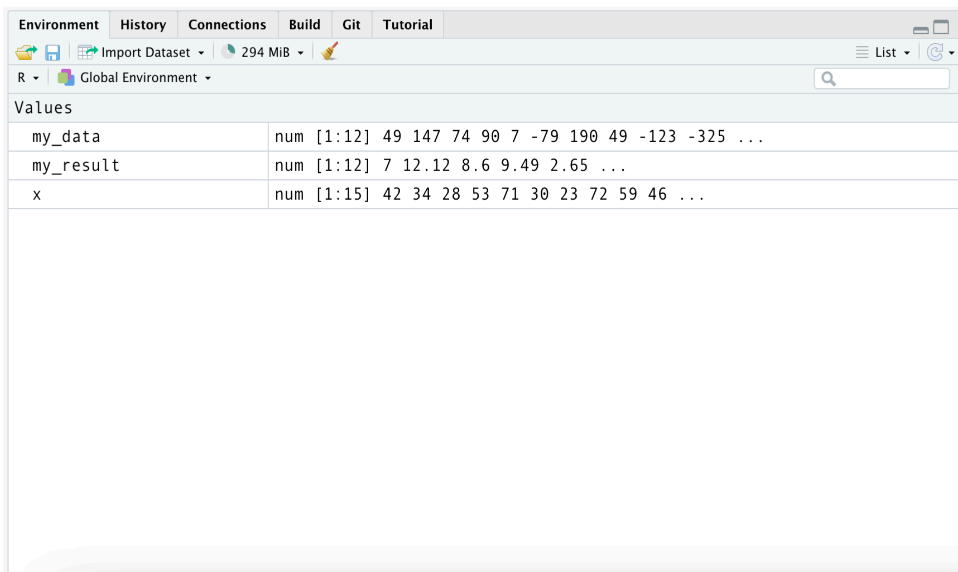
```
class(`<-`)
```

```
## [1] "function"
```

Data in R

Before we can do anything to a particular source of data (manipulate, analyze, visualize, model, etc.), we need to import in into the RStudio environment.

Right now we have three objects in our R environment (`my_data`, `my_result`, and `x`)



If we want to remove an object from the environment, we can use `rm()`.

```
rm(x)
```

To remove more than one object, separate them with commas.

```
rm(my_result, my_data)
```

Now we have an empty environment.

Loading data from packages

R comes 'out of the box' with a lot of data in the `datasets` package, which we can view using the command below:

```
library(help = "datasets")
```

To see the entire list of data–from datasets and any other R packages you've installed–use the `data()` function.

```
data()
```

Importing data

If we want to import a data from a local folder (like in our `Downloads` folder), we can use an import function like those from the `readr` package.

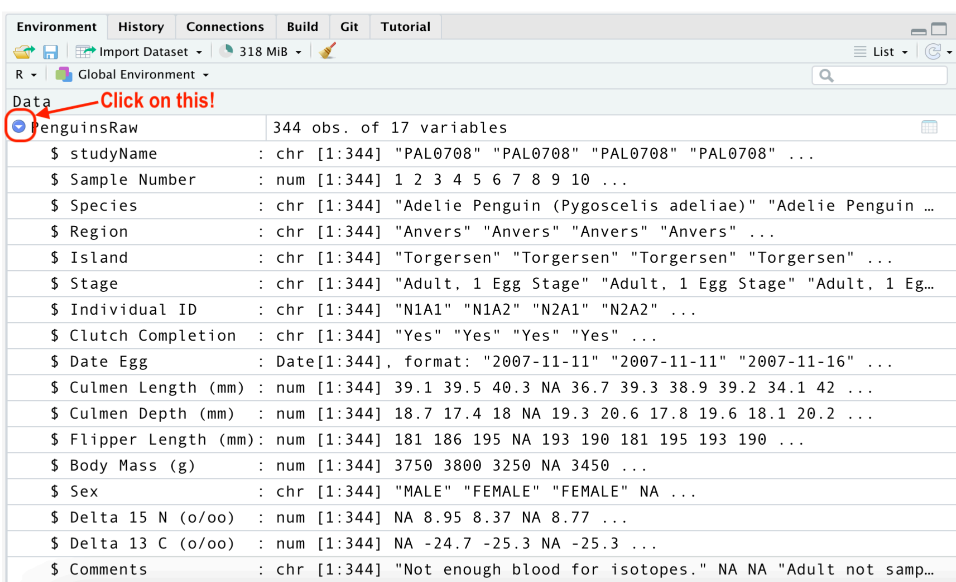
```
library(readr)
PenguinsRaw <- readr::read_csv(file = "~/Downloads/palmerpenguins.csv")
```

```
## Rows: 344 Columns: 17

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (9): studyName, Species, Region, Island, Stage, Individual ID, Clutch C...
## dbl  (7): Sample Number, Culmen Length (mm), Culmen Depth (mm), Flipper Leng...
## date (1): Date Egg

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

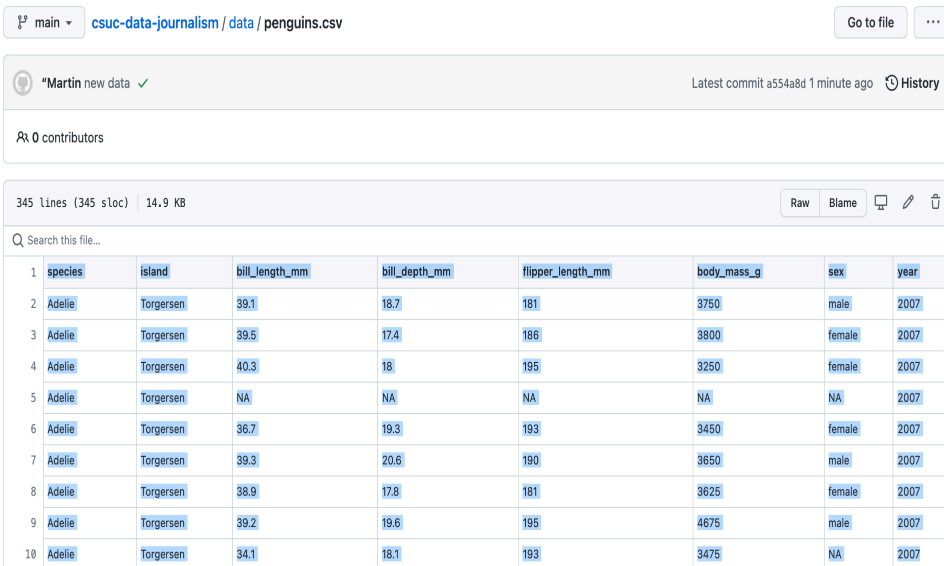
We can see `read_csv()` printed a lot of messages to the console to let us know a bit about how our .csv file was imported. We can also check the structure of our data by clicking on it in the **Environments** pane.



Other tricks for getting data

Another great trick for getting data into R comes from the `datapasta` package.

Let's head over to the penguins data, stored in the data/ folder in the Github repo for the course website. We are going to use our cursor and mouse to highlight the first ten rows of data (see the image below):



1	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
2	Adelie	Torgersen	39.1	18.7	181	3750	male	2007
3	Adelie	Torgersen	39.5	17.4	186	3800	female	2007
4	Adelie	Torgersen	40.3	18	195	3250	female	2007
5	Adelie	Torgersen	NA	NA	NA	NA	NA	2007
6	Adelie	Torgersen	36.7	19.3	193	3450	female	2007
7	Adelie	Torgersen	39.3	20.6	190	3650	male	2007
8	Adelie	Torgersen	38.9	17.8	181	3625	female	2007
9	Adelie	Torgersen	39.2	19.6	195	4675	male	2007
10	Adelie	Torgersen	34.1	18.1	193	3475	NA	2007

We're going to click `*cmd*` or `*ctrl*` + `*c*` to copy the data, then head back to RStudio.

In the console or a fresh .R script, we are going to enter the following

```
library(datapasta)
datapasta::tribble_paste()
```

And just like that, we see the following in the console:

```
tibble::tribble(
    ~species,    ~island, ~bill_length_mm, ~bill_depth_mm, ~flipper_length_mm, ~body_mass_g,     ~sex, ~year,
    "Adelie", "Torgersen",            39.1,           18.7,               181L,        3750L,   "male", 2007L,
    "Adelie", "Torgersen",            39.5,           17.4,               186L,        3800L, "female", 2007L,
    "Adelie", "Torgersen",            40.3,             18,               195L,        3250L, "female", 2007L,
    "Adelie", "Torgersen",              NA,             NA,                 NA,           NA,       NA, 2007L,
    "Adelie", "Torgersen",            36.7,           19.3,               193L,        3450L, "female", 2007L,
    "Adelie", "Torgersen",            39.3,           20.6,               190L,        3650L,   "male", 2007L,
    "Adelie", "Torgersen",            38.9,           17.8,               181L,        3625L, "female", 2007L,
    "Adelie", "Torgersen",            39.2,           19.6,               195L,        4675L,   "male", 2007L,
    "Adelie", "Torgersen",            34.1,           18.1,               193L,        3475L,       NA, 2007L
    )
```

Now all we need to do is add the assignment operator (`<-`) and the name for our new dataset (`penguin_sample`). We can print it to the console by just wrapping all the code in parentheses `()`

```r
(penguin_sample <- tibble::tribble(
     ~species,     ~island, ~bill_length_mm, ~bill_depth_mm, ~flipper_length_mm, ~body_mass_g,     ~sex, ~year,
     "Adelie", "Torgersen",            39.1,           18.7,               181L,        3750L,   "male", 2007L,
     "Adelie", "Torgersen",            39.5,           17.4,               186L,        3800L, "female", 2007L,
     "Adelie", "Torgersen",            40.3,             18,               195L,        3250L, "female", 2007L,
     "Adelie", "Torgersen",              NA,             NA,                 NA,           NA,       NA, 2007L,
     "Adelie", "Torgersen",            36.7,           19.3,               193L,        3450L, "female", 2007L,
     "Adelie", "Torgersen",            39.3,           20.6,               190L,        3650L,   "male", 2007L,
     "Adelie", "Torgersen",            38.9,           17.8,               181L,        3625L, "female", 2007L,
     "Adelie", "Torgersen",            39.2,           19.6,               195L,        4675L,   "male", 2007L,
     "Adelie", "Torgersen",            34.1,           18.1,               193L,        3475L,       NA, 2007L
     ))
```

If we're using R Markdown, we will see this below the code chunk:

A tibble: 9 x 8

species <chr>	island <chr>	bill_length_mm <dbl>	bill_depth_mm <dbl>	flipper_length_mm <int>	body_mass_g <int>	sex <chr>	year <int>
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007
Adelie	Torgersen	NA	NA	NA	NA	NA	2007
Adelie	Torgersen	36.7	19.3	193	3450	female	2007
Adelie	Torgersen	39.3	20.6	190	3650	male	2007
Adelie	Torgersen	38.9	17.8	181	3625	female	2007
Adelie	Torgersen	39.2	19.6	195	4675	male	2007
Adelie	Torgersen	34.1	18.1	193	3475	NA	2007

Conclusion

We've covered an introduction to R packages (base R and user-written packages), functions and objects, the 'store and explore' workflow, and how to load data into the R environment. Keep the document for future reference!